Available online at www.sciencedirect.com

SCIENCE DIRECT®

Genomics 87 (2006) 338 – 347

GENOMICS

www.elsevier.com/locate/ygeno

The presence of GC-AG introns in *Neurospora crassa* and other euascomycetes determined from analyses of complete genomes: Implications for automated gene prediction

Martijn Rep [a,*], Roselinde G.E. Duyvesteijn [a], Liane Gale [b], Thomas Usgaard [c], Ben J.C. Cornelissen [a], Li-Jun Ma [d], Todd J. Ward [c]

[a] *Plant Pathology, Swammerdam Institute for Life Sciences, Faculty of Science, University of Amsterdam, Kruislaan 318, 1098 SM Amsterdam, The Netherlands*
[b] *Cereal Disease Laboratory, Agricultural Research Service, United States Department of Agriculture, University of Minnesota, 1551 Lindig Street, St. Paul, MN 55108, USA*
[c] *Microbial Genomics and Bioprocessing Research Unit, Agricultural Research Service, United States Department of Agriculture, 1815 North University Street, Peoria, IL 61604, USA*
[d] *The Broad Institute of MIT and Harvard, 320 Charles Street, Cambridge, MA 02141-2023, USA*

Received 3 June 2005; accepted 24 November 2005
Available online 9 January 2006

Abstract

A combination of experimental and computational approaches was employed to identify introns with noncanonical GC-AG splice sites (GC-AG introns) within euascomycete genomes. Evaluation of 2335 cDNA-confirmed introns from *Neurospora crassa* revealed 27 such introns (1.2%). A similar frequency (1.0%) of GC-AG introns was identified in *Fusarium graminearum,* in which 3 of 292 cDNA-confirmed introns contained GC-AG splice sites. Computational analyses of the *N. crassa* genome using a GC-AG intron consensus sequence identified an additional 20 probable GC-AG introns in this fungus. For 8 of the 47 GC-AG introns identified in *N. crassa* a GC donor site is also present in a homolog from *Magnaporthe grisea, F. graminearum,* or *Aspergillus nidulans.* In most cases, however, homologs in these fungi contain a GT-AG intron or no intron at the corresponding position. These findings have important implications for fungal genome annotation, as the automated annotations of euascomycete genomes incorrectly identified intron boundaries for all of the confirmed and probable GC-AG introns reported here.

Keywords: Donor splice site; Noncanonical introns; Fungal genomes

Numerous fungal genome projects have recently been completed or are currently under way. After the landmark release of the genome sequence of *Neurospora crassa* [1], the first of a filamentous fungus, the genome sequences of the saprophytic ascomycetes *Aspergillus nidulans* and *Podospora anserina,* the mushroom *Coprinus cinereus,* the biotechnologically important fungi *Phanerochaete chrysosporium* and *Trichoderma reesei,* the plant pathogens *Magnaporthe grisea* [2], *Fusarium graminearum,* and *Ustilago maydis,* and the human pathogen *Cryptococcus neoformans* were made available to the public by the Broad Institute, the DOE Joint Genome Institute, and Genoscope. Many more genome sequencing projects involving filamentous fungi are currently under way.

Information from these projects is expected to advance medical, agricultural, and biotechnological research. However, the vast majority of protein-coding genes within these genomes have not been experimentally characterized, making accurate methods for automated gene prediction essential. Determining the correct exon boundaries is a critical problem for gene prediction based on genomic sequences [3]. For small introns, which constitute a separate class of introns with a narrow length distribution [4,5], short sequence motifs contain enough information to predict the correct intron/exon boundaries in 85–95% of cases, depending on the organism [5]. However, assuming an average of two introns per gene, the

[*] Corresponding author. Fax: +31 20 5257934.
E-mail address: m.rep@uva.nl (M. Rep).

0888-7543/$ - see front matter © 2005 Elsevier Inc. All rights reserved.
doi:10.1016/j.ygeno.2005.11.014

intron–exon structure of at best 9 of 10 genes will be correctly predicted. Even in the well-characterized yeast *Saccharomyces cerevisiae*, only 61 of 87 intron predictions were found to be correct [6].

Spliceosomal introns generally begin with GT and end with AG dinucleotide motifs that are referred to as donor and acceptor splice sites, respectively. However, introns with noncanonical splice sites have been identified and have the potential to confound accurate gene prediction further [7]. Therefore, for automated gene annotation based on a genome sequence it is important to establish if an organism or a group of organisms has alternative intron isoforms and to estimate the frequency of noncanonical intron splice site motifs within a given genome. Based on comparisons of cDNA and genomic sequences in mammals, over 90% of noncanonical introns have GC-AG splice sites [7]. In addition, the few noncanonical introns reported previously for yeast [6,8] and the single one from a filamentous fungus [9] have GC-AG splice sites, indicating that this isoform is likely the most important for accurate gene prediction. To estimate the frequency of GC-AG introns in euascomycetes and to assess their impact on current genome annotations 2335 cDNA-confirmed introns from *N. crassa* were examined for noncanonical intron splice sites. Based on these sequences, a GC-AG splice consensus was developed to predict additional GC-AG introns in the *N. crassa* genome. In addition, the phylogenetic distribution of GC-AG introns identified in the *N. crassa* genome was examined by comparative analyses of homologous sequences in *A. nidulans, F. graminearum*, and *M. grisea*, and the existence of GC-AG introns in two *Fusarium* species was verified experimentally. The results indicate that automated annotations of fungal genomes can be substantially improved by consideration of GC-AG introns.

Results and discussion

Identification of 27 GC-AG introns in N. crassa

To determine whether alternatives to the standard GT-AG intron isoform were present within the *N. crassa* genome, 29,625 ESTs were aligned to Release 3 of the *N. crassa* genome sequence at the Broad Institute using the sequence alignment tool BLAT [10]. Of these, 24,746 could be aligned with at least 99% sequence identity, with 10,124 spanning one or more apparent introns. From this set, 2335 unique introns were derived. Twenty-seven introns possessed GC donor sites (Table 1), while all of the other identified introns had the standard GT-AG configuration. All 27 GC-AG introns were manually verified. This frequency of GC-AG introns (1.2%) is somewhat higher than the frequencies found in *Caenorhabditis elegans* (0.6%) [11] and mammals (0.7%) [9]. The current models for genes containing GC-AG introns differ in various ways from the gene models that are based on the presence of a GT-AG intron (Table 1). In 14 cases, an overlapping GT-AG intron is annotated with the GT donor site upstream or downstream of the GC donor site. In two of these cases the acceptor site is also incorrectly predicted in the current annotation. In 7 cases the intron was missed altogether. The remaining 6 GC-AG introns were found outside current gene models, which underscores the importance of correct intron definition for gene prediction.

In silico prediction of GC-AG introns

Having confirmed the existence of several GC-AG introns in *N. crassa*, we asked whether we could find additional GC-AG introns in *N. crassa* using an in silico approach. We first designed a consensus GC-AG intron sequence (G/GCAAGT N{30,70} CTAAC N{6,20} YAG) based on the 27 EST-confirmed introns listed in Table 1. Only 5 of the 27 introns fully conform to this consensus. However, the purpose here was not to be exhaustive but to explore the potential of an in silico approach. By using only the most common bases at some positions (especially at donor and branch sites) and restriction of the distances between donor and branch sites and between branch and acceptor sites, we aimed to reduce the number of false positives. In the genome of *N. crassa*, 72 sites match our GC-AG consensus pattern. However, nonintron patterns of similar complexity are present at comparable frequencies (not shown). Therefore, to assess which of the sites could be real introns, flanking sequences (putative exons) were translated and the products compared to proteins in public databases. Predicted GC-AG introns were considered highly probable if the level of protein sequence identity in the relevant part of the proteins (encoded by the neighboring exons of the candidate intron) allowed unequivocal alignment with proteins found in public databases. With these criteria, 24 of the 72 potential introns were considered highly probable. Four of these were already identified with the EST/genome comparison described above (in NCU01417.1, NCU02207.1, and NCU03195.1 and an unrecognized gene in contig 3.458, Table 1); the remaining 20 are listed in Table 2. Based on alignments with homologs, 2 of the 72 potential introns were considered to be false. In these 2 cases (in NCU06143.1 and NCU07919.1), introns are currently annotated with the same acceptor sites but with GT donor sites downstream of the proposed GC donor sites (24 and 8 bp, respectively) that are more likely based on amino acid alignments of the translation products. The remaining 46 potential introns did not reside in genes with close homologs. Among these was 1 intron that was identified with the EST/genome comparison (in NCU08751.1, Table 1). Probably, there are more true introns among the 46 potential introns without close homologs in sequence databases.

Donor and branch sites in GC-AG introns appear to be more conserved than those of GT-AG introns

It is remarkable that the pattern used for in silico detection of GC-AG introns detects 5 of the 27 GC-AG introns (19%) found with the EST/genome comparison, while the GT version of the pattern (which differs only in the donor site) detects only 333 (2%) of the estimated ~17,000 GT-AG introns in the genome. This cannot be attributed to close phylogenetic relatedness between the GC-AG introns because there is no

Table 1
N. crassa GC-AG introns found by EST–genome comparison

Contig	Intron pos.	Gene	Donor site[a]	S1[b]	Branch site	S2[c]	Current gene model	*M. grisea*[d]	*F. graminearum*[d]	*A. nidulans*[d]
3,13	75527 –75598	NCU00385.1	G/GCATGT	39	CTCTAAC	18	Intron 1: GT 755 bp upstream; incorrect start; 102 bp missing from exon 1 and 24 bp added to exon 1	MG05742.4: GT-AG intron (GT 454 bp upstream in current annotation); intron placement and AA seq supports GC-AG annotation in Nc	FG06648.1: unrecognized GC-AG intron; GT 12 bp upstream in current annotation	AN0277.2: No intron; AA seq supports GC-AG annotation in Nc
3,29	60389–60497	NCU00828.1	G/GCAAGC	66	GGCTTAC	28	Intron 4 (GC-AG) missed: Intron 3 extends through exon4 (144 bp) and intron 4 (109 bp)	MG10836.4: truncated due to end of contig	FG00622.1: GT-AG intron; intron placement and AA seq supports GC-AG annotation in Nc	AN1483.2: No intron; AA seq supports GC-AG annotation in Nc
3,32	93997–94061	NCU00889.1	A/GCAGGT	35	ATCTAAC	15	Intron 1: GT 4 bp downstream, GC-AG intron placement not supported by alignments	MG01179.4; GT-AG intron 4 bp downstream	FG00808.1; GT-AG intron 4 bp downstream	AN9072.2; GT-AG intron 4 bp downstream
3,38	151188–151288	None called	T/GCAAGT	67	GACTAAT	19		MG10665.4: too divergent	FG00133.1: too divergent	No similarity
3,51	124690–124760	NCU01328.1	G/GCAAGC	41	TGCTAAC	15	Intron 5: GT 18 bp downstream	MG02471.4: GT-AG intron; intron placement and AA seq supports GC-AG annotation in Nc	FG09998.1: unrecognized GC-AG intron; GT 9 bp upstream in current annotation	AN0688.2: No intron, ; AA seq supports GC-AG annotation in Nc and Fg
3,57	3588–3652	NCU01417.1	G/GCAAGT	40	CTCTAAC	10	Intron 5: GT 12 bp uptream	MG00689.4: GT-AG intron; intron placement and AA seq supports GC-AG annotation in Nc	FG10126.1: GT-AG intron; intron placement and AA seq supports GC-AG annotation in Nc	AN5833.2: GT-AG intron; intron placement and AA seq supports GC-AG annotation in Nc
3.64	113190–113250	NCU01535.1	T/GCAAGT	34	TGCTAAC	12	Intron 3: GT 14 bp downstream	MG01081.4: GT-AG intron; intron placement and AA seq supports GC-AG annotation in Nc	FG04329.1: GT-AG intron; intron placement and AA seq supports GC-AG annotation in Nc	AN6258.2: No intron, AA seq too divergent
3,104	20742–20816	NCU02207.1	G/GCAAGT	41	AACTAAC	19	Intron 1: GT 677 bp upstream; incorrect start; 3 bp missing from exon 1 and 27 bp added to exon 1	MG11979.4: GT-AG intron; intron placement and AA seq supports GC-AG annotation in Nc	FG05778.1: GT-AG intron; intron placement and AA seq supports GC-AG annotation in Nc	AN0381.2: GT-AG intron; intron placement and AA seq supports GC-AG annotation in Nc

3,119	12077 –12134	None called	G/GCAAGC	36	TGCTAAT	7		MG03576.4: GT-AG intron; intron placement and AA seq supports GC-AG annotation in Nc	FG11628.1: GT-AG intron; intron placement and AA seq supports GC-AG annotation in Nc	AN3467.2: GT-AG intron; intron placement and AA seq supports GC-AG annotation in Nc
3,152	65794 –65862	None called	A/GCAAGT	35	TGCTAAC	17		No similarity	No similarity	No similarity
3,165	62085 –62197	NCU03151.1	G/GCAAGT	77	TACTAAC	21	Intron 1: GT 84 bp upstream	MG02710.4: GT-AG intron; intron placement and AA seq supports GC-AG annotation in Nc	FG08677.1:GT-AG intron; intron placement and AA seq supports GC-AG annotation in Nc	AN8692.2: No intron; AA seq supports GC-AG annotation in Nc
3,167	29779 –29865	NCU03195.1	G/GCAAGT	63	ATCTAAC	9	Intron 2 (GC-AG) missed: Intron 1 extends through exon 2 (79 bp), intron 2 (87 bp), and 80 bp of exon3	MG01257.4: unrecognized GC-AG intron; GT 267 bp upstream and AG 139 bp downstream in current annotation	FG01258.1:GT-AG intron; intron placement and AA seq supports GC-AG annotation in Nc	No similarity
3,167	153693– 153818	NCU03233.1	A/GCAAGT	102	AGCTAAC	9	Intron 2 (GC-AG) missed: protein has 42 extra AA	MG10604.4: GT-AG intron; intron placement and AA seq supports GC-AG annotation in Nc	FG01195.1: too divergent	AN3922.2: too divergent
3,201	179474– 179539	NCU03717.1	G/GCATGT	36	TTCTAAC	15	Intron 1 (GC-AG) missed: protein has 22 extra AA	MG04647.4: No intron	FG07146.1: No intron; AA seq supports GC-AG annotation in Nc	AN1080.2: No intron; AA seq supports GC-AG annotation in Nc
3,202	64335 –64414	NCU03766.1	G/GCACGT	47	TGCTGAC	18	Intron 2: GT 21 bp downstream	MG02462.4: GT-AG intron; intron placement and AA seq supports GC-AG annotation in Nc	FG06819.1: No intron; AA seq supports annotation in Nc	AN1080.2: No intron; AA seq supports GC-AG annotation in NC
3,221	21162–21224	NCU04237.1	A/GCAAGT	32	GACTGAC	16	Intron 2 (GC-AG) missed: Intron 1 starts 38 bp downstream of real donor site and extends through exon 2 (29 bp) and intron 2 (63 bp)	MG09544.4: No intron	FG04568.1: No intron; AA seq supports GC-AG annotation in Nc	AN4305.2: No intron; AA seq supports GC-AG annotation in Nc
3.230	32805 –32934	NCU04467.1	G/GCAAGT	101	GACTAAC	14	Intron 1: GT 18 bp downstream	No similarity	No similarity	No similarity
3,273	88063 –88121	NCU04928.1	T/GCAAGT	35	TACTAAC	9	Intron 1: GT 13 bp upstream and AG 7 bp upstream	MG09381.4: too divergent	No similarity	AN7265.2: too divergent
3,301	18648 –18718	NCU05291.1	G/GCGAGT	44	GACTAAC	12	Intron 1: GT 28 bp downstream and AG 55 bp downstream	MG07134.4: GT-AG intron; intron placement and AA seq supports GC-AG annotation in Nc	FG06163.1: GT-AG intron; intron placement and AA seq supports GC-AG annotation in Nc	AN5866.2: unrecognized GC-AG intron; GT 18 bp upstream in current annotation
3,305	30284 –30346	None called	G/GCACGT	38	TCCTAAC	10		No similarity	No similarity	No similarity

(continued on next page)

Table 1 (*continued*)

Contig	Intron pos.	Gene	Donor site[a]	S1[b]	Branch site	S2[c]	Current gene model	*M. grisea*[d]	*F. graminearum*[d]	*A. nidulans*[d]
3,356	61097–61178	NCU06110.1	C/GCAAGT	53	TACTAAC	14	Intron 1: first of 2 introns in mRNA leader	MG03098.4	FG02469.1	AN3928.2
3,359	61963–62034	None called	G/GCAGGT	46	TGCTAAC	11		No similarity	No similarity	No similarity
3,371	131629– 131773	NCU06451.1	G/GCATGT	115	TGCTAAC	15	Intron 1: GT 21 bp downstream	MG06467.4: GT-AG intron; intron placement and AA seq supports GC-AG annotation in Nc	FG07286.1: GT-AG intron; intron placement and AA seq supports GC-AG annotation in Nc	No similarity
3.458	91647–91723	None called	G/GCAAGT	44	TGCTAAC	18		MG01508.4: unrecognized GC-AG intron; ORF ends in this intron in current annotation	FG01940.1: GT-AG intron (not recognized in current annotation); intron placement and AA seq supports GC-AG annotation in Nc	AN5869.2: GT-AG intron; intron placement and AA seq supports GC-AG annotation in Nc
3,517	89482–89549	NCU08554.1	C/GCAAGT	32	AGCTGAC	21	Intron 1: GT 600 bp upstream; incorrect start and 2 AA incorrectly deleted from exon1	MG07389.4: GT-AG intron; intron placement and AA seq supports GC-AG annotation in Nc	FG06532.1: GT-AG intron; intron placement and AA seq supports GC-AG annotation in Nc	AN6145.2: No intron; AA seq supports GC-AG annotation in Nc
3,538	19652–19748	NCU08691.1	G/GCGAGT	59 or 65	AACTAAC or CTCTAAC	23 or 17	Intron 1: GT 423 bp upstream; incorrect start; 60 bp missing from exon 1 and 97 bp added to exon 1	MG04173.4: GT-AG intron; intron placement and AA seq supports GC-AG annotation in Nc	FG10545.1: GT-AG intron; intron placement and AA seq supports GC-AG annotation in Nc	AN8965.2: No intron; AA seq supports GC-AG annotation in Nc
3,542	2022–2092	NCU08751.1	G/GCAAGT	37	TGCTAAC	19	Intron 1 (GC-AG) missed: protein has 8 extra AA at N-terminus	No similarity	No similarity	No similarity

[a] Underlined: divergence from consensus (G/GCAAGT N{30,70} CTAAC N{6,20} YAG).

[b] S1: distance between donor and branch sites.

[c] S2: distance between branch and acceptor (YAG) sites.

[d] Closest homolog in respective species; in bold: GC-AG intron in the same position as in *N. crassa*.

sequence similarity beyond the splice signals and there are no paralogs among the genes that they reside in. Also, the median length of the EST-confirmed GT-AG introns is similar to the that of the GC-AG introns (77 versus 72). Together, these observations imply that GC-AG introns exhibit a higher level of similarity to "optimal" splice signals. Indeed, the donor and branch sites of 26 of the 27 confirmed GC-AG introns differ in at most one position from the pattern (even when including an extra purine in the branch site (*R*CTAAC), which was not part of the search pattern but suggested by the sequences of all 47 introns in Tables 1 and 2). The only exception is the intron in NCU00889.1, which deviates at two positions from the donor site pattern. However, this intron may have been aberrantly or alternatively spliced, as discussed below. These observations are in agreement with reports on mammalian GC-AG introns, which also appear to tolerate less variability in sequence, especially around the donor site [7,12,13].

Comparative analyses across genomes of euascomycetes

To estimate the extent of conservation of GC splice donor sites, we searched for the closest homologs of the 47 *N. crassa* genes containing confirmed or probable GC-AG introns in the euascomycetes *F. graminearum, M. grisea*, and *A. nidulans*. Analysis of these homologs revealed that, in most cases, either a GT-AG intron is present at exactly the same position as the GC-AG intron in *N. crassa* or no intron is present (Table 1). In 8 cases, a GC-AG intron was present at the same position in an *N. crassa* gene and a homologous gene of another species (in NCU00385.1, encoding an ATP synthase δ chain; NCU01328.1, encoding a probable transketolase; NCU03195.1, encoding a potential tRNA dihydrouridine synthase; NCU05291.1, encoding a potential polyamine *N*-acetyl transferase; an unrecognized gene in contig 3.458; NCU01768.1; NCU06729.1, encoding the G-protein α subunit Gna2; and NCU07554.1, encoding a chromosome scaffold protein). No intron is conserved in more than two genera, and we consider it unlikely that the level of conservation that we observe could be related to regulation of (alternative) splicing. Indeed, there are no ESTs corresponding to alternatively spliced or unspliced RNAs. Also in human and *C. elegans* the majority of GC-AG introns appears to be constitutively spliced [11,13]. In a recent paper describing the analysis of a large number of ESTs of *M. grisea,* the single EST corresponding to the use of a GC donor site represented a rare splice event (1 of 66 transcripts from a single gene). It is unclear, however, whether this was related to gene regulation (leading to a product with different properties) or just a case of missplicing [14].

One remarkable case listed in Table 1 could also be the result of a missplicing event. In NCU00889.1 (encoding a Ras family member), a GC donor site is implied in the first intron by EST NCSM4F3T3 (subtracted mycelial *N. crassa* cDNA clone SM4F3) (Table 1, contig 3.32). However, the currently annotated GT donor site 4 bp downstream of the GC donor site is the one that leads to the correct translation product based on comparison with homologs in other fungi. In *F. graminearum, M. grisea,* and *A. nidulans* there is a GT-AG intron at the same (+4) position. EST NCSM4F3T3 could therefore be the result of aberrant splicing, but it remains unclear why there is no EST corresponding to the use of the GT donor site.

To obtain experimental evidence for the existence of GC-AG introns in euascomycetes other than *N. crassa,* we also performed an EST/genome comparison for *F. graminearum.* Intron position and sequence determinations were made for 292 loci based on assessment of positional homology between previously published expressed sequence tags [15] and genomic sequences from the *F. graminearum* (PH-1, NRRL 31084) genome sequence database (http://www.broad.mit.edu/annotation/fungi/fusarium/). From these analyses, the presence of three GC-AG introns (1.0%) could be inferred (in FG01085.1, FG06370.1, and FG06931.1). In addition, the presence of a GC-AG intron in the *F. oxysporum* gene for subunit c of the V-type ATPase (GenBank Accession No. AY587846) was confirmed with cDNA sequencing. Its ortholog in *F. graminearum* (FG01328.1) also contains a GC-AG intron at that position, while those of *M. grisea* (MG06349.4) and *N. crassa* (NCU09897.1) do not (the latter contain, respectively, a GT-AG intron and no intron at the corresponding position). The DNA sequences of *F. graminearum* GC-AG introns were verified with independent genomic sequence data for PH-1 (genome sequence strain) and a second strain of *F. graminearum* (NRRL 34097). Evolutionary conservation of these noncanonical intron motifs was assessed by comparison with sequences from closely related fusaria: *F. asiaticum* [16], *F. lunulosporum, F. cerealis, F. culmorum, F. pseudograminearum,* and *F. sporotrichioides.* Interestingly, the GC-AG motif in the serine phosphatidyltransferase (encoded by FG06370.1) appears to be a recent mutation restricted to the *F. graminearum* species complex [16], because this GC-AG was also found in *F. asiaticum,* but *F. culmorum, F. cerealis,* and *F. lunulosporum* had GT-AG borders. For the other two genes, the GC-AG border was found in all species examined, indicating that the mutation is at least as old as the trichothecene-producing clade of *Fusarium*.

Toward an automated recognition of GC-AG introns

Since the GC-AG intron frequency in *N. crassa* is about 1.2%, and the total number of predicted introns in this fungus is about 17,000, the total number of GC-AG introns in this fungus is expected to be around 200. With our strict consensus pattern, already an additional 20 probable GC-AG introns were found, with several more likely to be among the potential introns that could not be confirmed by alignment to homologs in other euascomycetes. Among the GC-AG introns not identified in this study is the only such intron that was previously reported for *N. crassa,* in the *qa* repressor gene (donor site G/GCACGT, branch site TACTAAC) [9]. The existence of GC-AG introns in the fungal kingdom has not yet been widely recognized, but has important consequences for (automated) gene annotation. The in silico approach for GC-AG intron detection described here was used for an initial survey only, but elements thereof may be integrated into existing gene prediction programs such as FGENESH [17] and

Table 2
Probable *N. crassa* GC-AG introns found by *in silico* genome survey

Contig	Intron pos.	Gene	Donor site[a]	S1[b]	Branch site	S2[c]	Current gene model	*M. grisea*[d]	*F. graminearum*[d]	*A. nidulans*[d]
3,11	65185–65243	NCU00217.1	G/GCAAGT	32	AGCTAAC	12	Intron 4: GT 72 bp upstream	MG06279.4: GT-AG intron; intron placement and AA seq supports GC-AG annotation in Nc	FG01293.1: No intron; AA seq supports GC-AG annotation in Nc	AN3650.2: GT-AG intron; intron placement and AA seq supports GC-AG annotation in Nc
3,53	48114–48167	NCU01382.1	G/GCAAGT	28	TGCTAAC	11	ORF starts 116 bp downstream of GC-AG intron (upstream exons missed)	No similarity	FG06086.1: No intron; AA seq supports GC-AG annotation in Nc	AN2391.2: No intron; AA seq supports GC-AG annotation in Nc
3,72	49560–49629	NCU01654.1	G/GCAAGT	47	TGCTAAC	8	Intron 1: GT 38 bp upstream	MG07197.4: No intron; AA seq supports GC-AG annotation in Nc	FG01419.1: GT-AG intron; intron placement and AA seq supports GC-AG annotation in Nc	AN8280.2: GT-AG intron; intron placement and AA seq supports GC-AG annotation in Nc
3,75	109136–109230	NCU01768.1	G/GCAAGT	64	AACTAAC	16	Intron 1: GT 19 bp downstream and AG 16 bp downstream (correct gene model is AL355926)	No similarity	FG00299.1: unrecognized **GC-AG intron**; GT 33 bp upstream in current annotation	AN6286.2: GT-AG intron; intron placement and AA seq supports GC-AG annotation in Nc
3,114	3401–3466	NCU02382.1	G/GCAAGT	41	GGCTAAC	10	Intron 1: GT 209 bp upstream; incorrect start	MG03513.4: too divergent	FG06362.1: GT-AG intron; intron placement and AA seq supports GC-AG annotation in Nc	AN0927.2: No intron; AA seq supports GC-AG annotation in Nc
3,143	58805–58860	NCU02777.1	G/GCAAGT	30	CACTAAC	11	Intron 1: GT 78 bp upstream	MG01613: GT-AG (but shifted with respect to annotated intron: GT 4 bp upstream, AG 23 bp downstream); intron placement and AA seq supports GC-AG annotation in Nc	FG01107.1: GT-AG intron; intron placement and AA seq supports GC-AG annotation in Nc	AN1637.2: GT-AG intron; intron placement and AA seq supports GC-AG annotation in Nc
3,150	46105–46161	NCU02853.1	G/GCAAGT	35	AACTAAC	7	Intron 1: GT 4 bp downstream and AG 97 bp downstream	MG04865.4: too divergent	FG11388.1: GT-AG intron; intron placement and AA seq supports GC-AG annotation in Nc	AN8357.2: GT-AG intron; intron placement and AA seq supports GC-AG annotation in Nc
3,161	63123–63201	NCU03087.1	G/GCAAGT	50	GGCTAAC	14	Intron 1: GT 49 bp upstream and AG 8 bp downstream	MG01272.4: GT-AG intron; intron placement and AA seq supports GC-AG annotation in Nc	FG01222.1: GT-AG intron; intron placement and AA seq supports GC-AG annotation in Nc	AN1141.2: GT-AG intron; intron placement and AA seq supports GC-AG annotation in Nc
3,164	102409–102479	NCU03124.1	G/GCAAGT	37	AGCTAAC	19	Intron 2: GT 27 bp upstream (correct gene model is AF494376 (Yang et al. 2002))	MG03696.4: No intron; AA seq supports GC-AG annotation in Nc	FG00677.1: GT-AG intron; intron placement and AA seq supports GC-AG annotation in Nc	AN1485.2: GT-AG intron; intron placement and AA seq supports GC-AG annotation in Nc
3,214	229296–229366	NCU04059.1	G/GCAAGT	41	TGCTAAC	15	Intron 1: GT 21 bp upstream and AG 21 bp downstream	MG00594.4: GT-AG intron; intron placement and AA seq supports GC-AG annotation in Nc	FG05337.1: GT-AG intron; intron placement and AA seq supports GC-AG annotation in Nc	AN5499.2: No intron; AA seq supports GC-AG annotation in Nc

3,311	117403–117478	None called	G/GCAAGT	41	CGCTAAC	20		MG00437.4: GT-AG intron; intron placement and AA seq supports GC-AG annotation in Nc	FG08328.1: GT-AG intron; intron placement and AA seq supports GC-AG annotation in Nc	Contig 1.104 (58305-57846): GT-AG intron; intron placement and AA seq supports GC-AG annotation in Nc
3,312	231277–231354	NCU05608.1	G/GCAAGT	49	TGCTAAC	14	Intron 1 (GC-AG) missed (in-frame)	Contig 2.1040 (33357-end of contig): too divergent	FG08133.1: No intron; AA seq supports GC-AG annotation in Nc	AN5982.2: GT-AG intron; intron placement and AA seq supports GC-AG annotation in Nc
3,354	515–575	NCU06080.1	G/GCAAGT	32	AACTAAC	14	Intron 1: GT 10 bp upstream and AG 14 bp downstream	MG04975.4: No intron; AA seq supports GC-AG annotation in Nc	FG09186.1: No intron; AA seq supports GC-AG annotation in Nc	AN5354.2: No intron; AA seq supports GC-AG annotation in Nc (current annotation suggests a GT-AG intron starting 4 bp downstream)
3,389	6684–6744	NCU06729.1	G/GCAAGT	34	AACTAAC	12	Intron 2: GT 33 bp downstream and AG 72 bp downstream (correct gene model is AF004846 (Baasiri 1997))	MG04204.4: unrecognized **GC-AG intron** at same position; GT 4 bp downstream, AG 70 bp downstream in current annotation	FG09988.1: GT-AG intron; intron placement and AA seq supports GC-AG annotation in Nc	AN0651.2: too divergent
3,429	29279–29343	NCU07375.1	G/GCAAGT	37	AGCTAAC	13	Intron 1: GT 54 bp upstream	MG00346.4: No intron; AA seq supports GC-AG annotation in Nc	FG01311.1: No intron; AA seq supports GC-AG annotation in Nc	AN5935.2: No intron; AA seq supports GC-AG annotation in Nc
3,445	5439–5487	NCU07554.1	G/GCAAGT	33	TACTAAC	14	Intron 2: GT 25 bp upstream, AG 11 bp downstream	MG04988: GT-AG intron; intron placement and AA seq supports GC-AG annotation in Nc	FG06754.1: unrecognized **GC-AG intron**; GT 54 bp downtream, AG 48 bp downstream in current annotation	AN6364.2: No intron; AA seq supports GC-AG annotation in Nc
3,550	45053–45107	NCU08852.1	G/GCAAGT	31	GACTAAC	9	Intron 2 (GC-AG) missed, and an unlikely intron just downstream	MG08613: No intron; AA seq supports GC-AG annotation in Nc	FG05924.1: No intron; AA seq supports GC-AG annotation in Nc	AN3129.2: No intron; AA seq supports GC-AG annotation in Nc
3,562	102534–102592	NCU09006.1	G/GCAAGT	35	AACTAAC	9	Intron 1: GT 39 bp upstream	MG01669.4: GT-AG intron; intron placement and AA seq supports GC-AG intron in Nc	FG05561.1: GT-AG intron; similarity of upstream exon to Nc too low to compare intron placement	AN2298.2; GT-AG intron; intron placement and AA seq supports GC-AG intron in Nc
3,568	3775–3858	NCU09070.1	G/GCAAGT	51	TACTAAC	18	Intron 1 (GC-AG) missed (in-frame)	MG02723.4: GT-AG intron; intron placement and AA seq supports GC-AG intron in Nc	FG08801.1: GT-AG intron; intron placement and AA seq supports GC-AG annotation in Nc	AN8724.2: No intron; AA seq supports GC-AG annotation in Nc
3,667	6501–6592	NCU09817.1	G/GCAAGT	60	CGCTAAC	17	Intron 3: GT donor 39 bp upstream	MG05734.4; no intron; AA seq supports GC-AG annotation in Nc	FG00478.1: GT-AG intron; intron placement and AA seq supports GC-AG xintron in Nc	AN0354.2; GT-AG intron; intron placement and AA seq supports GC-AG intron in Ncl

[a] Introns were found with the pattern G/GCAAGT N{30,70} CTAAC N{6,20} YAG.

[b] S1: distance between donor and branch sites.

[c] S2: distance between branch and acceptor (YAG) sites.

[d] Closest homolog in respective species; in bold: GC-AG intron in the same position as in *N. crassa.*

GENSCAN [18] or further developed into splice site probability models for prediction of noncanonical introns. The novelty of such a procedure would be that it would start with detection of potential introns in the whole genome.

Based on the analysis reported here, further steps can be taken in validating automated annotation of GC-AG intron-containing genes in a conservative way:

1. Combine neighboring "exons" and BLAST search with the predicted protein sequences (three frames) against predicted proteins of a number of fungal genomes (at least three). Phylogenetic distance should be such that there is a fair chance of conservation of protein sequences as well as intron positions so that these can be used for verification purposes (see steps below). Comparison of several euascomycete genomes as was done in this study appears to work well.
2. Discard the GC-AG intron-containing gene if there is no BLAST hit (i.e., no in silico verification is possible).
3. If there are BLAST hits, proceed if at least one of the protein sequence alignments includes the position of the putative intron. In the analysis reported here, the intron position was marked in the protein sequence with an inserted "X". Therefore, the BLAST program needed to introduce a gap of one amino acid in the target protein and still align (part of) the upstream and downstream sequences.
4. See if in the gene model for any of the homologous proteins there is a (GT-AG) intron at the same position. If so, accept the GC-AG intron as probable.

In the present study, 4 of the 20 introns found with the pattern search would be rejected by the criterion of intron position conservation (Table 2). One could introduce alternative criteria for verification of potential introns. Of the four *N. crassa* introns without conserved intron position in any of the homologs, three (in NCU01382.1, NCU07375.1, and NCU08852.1) fulfill the following criteria: (1) significance of BLAST hit is at least e^{-30} and (2) the corresponding alignment extends over at least 70 residues upstream as well as downstream of the intron position. With these criteria, the following step would be:

5. If no conservation of intron position is found, check if at least one of the BLAST hits was significant at e^{-30} or better and if the alignment extends over at least 70 residues upstream as well as downstream of the intron position.

One GC-AG intron without conserved intron position (in NCU06080.1) does not pass these criteria because the upstream alignment extends over only eight residues. It was still considered likely to be a true intron because six of these eight residues are identical and amino-terminal in all homologs, resulting in coinciding translational starts.

Of course, a number of true introns will be discarded using this procedure for several reasons: (1) There is no homologous protein predicted to be encoded by the genomes used for comparisons; (2) homologs are found, but the position of the intron is not conserved or homology on one side of the intron position is too low to yield a BLAST alignment; (3) intron position is not conserved and the protein alignment does not pass the criteria mentioned in step 5; and (4) there are errors in the gene model(s) of the homolog(s) such that intron positions appear not to be conserved (note several cases in Tables 1 and 2 in which adjustments were made in gene models to improve protein alignments leading to conserved intron positions—in some cases the error was due to an unrecognized GC-AG intron). As an indication of the frequency of false negatives that may be expected, 10 of the 27 experimentally confirmed introns listed in Table 1 would not be verified using this procedure.

The usefulness of this procedure extends beyond identification of GC-AG introns. With modifications, classical GT-AG introns could also be identified, complementing current methods of gene model construction and possibly leading to discovery of previously unrecognized genes, such as the *N. crassa* gene containing a GC-AG intron in contig 3.311 (Table 2).

Materials and methods

DNA sequencing

DNA sequencing was performed with ABI BigDye chemistry version 3.0 and an ABI 3730 genetic analyzer (Applied Biosystems, Foster City, CA, USA) as previously described [19].

In silico intron searches

Fungal genome sequences were downloaded from the Broad Institute Web site (www.broad.mit.edu) and analyzed with home-made PERL scripts using MacPerl (http://www.ptf.com/macperl/). Briefly, all sites in the *N. crassa* genome sequences corresponding to the GC-AG intron consensus sequence [G/GCAAGT N{30,70} CTAAC N{6,20} YAG] were extracted (72 sites). In addition, the frequency of sites matching the canonical (GT donor splice site) derivative thereof [G/GTAAGT N{30,70} CTAAC N{6,20} YAG] was determined (333 sites). From all sites extracted with the GC-AG intron pattern, flanking sequences (up to 900 bases on each side) were combined. The longest ORF that overlapped with the presumed intron was translated and the product was used to search for homologous sequences in public databases at NCBI (http://www.ncbi.nlm.nih.gov/BLAST/). Alignments were inspected manually to judge whether the presumed intron was likely to be real (see text for details).

Acknowledgments

The research of M. Rep has been made possible by a fellowship of the Royal Netherlands Academy of Arts and Sciences. R.G.E. Duyvesteijn was supported by the Utopa Foundation.

References

[1] J.E. Galagan, et al., The genome sequence of the filamentous fungus *Neurospora crassa*, Nature 422 (2003) 859–868.

[2] R.A. Dean, et al., The genome sequence of the rice blast fungus *Magnaporthe grisea*, Nature 434 (2005) 980–986.

[3] J. Wang, et al., Vertebrate gene predictions and the problem of large genes, Nat. Rev. Genet. 4 (2003) 741–749.

[4] E.V. Kriventseva, M.S. Gelfand, Statistical analysis of the exon–intron structure of higher and lower eukaryote genes, J. Biomol. Struct. Dyn. 17 (1999) 281–288.

[5] L.P. Lim, C.B. Burge, A computational analysis of sequence features

involved in recognition of short introns, Proc. Natl. Acad. Sci. USA 98 (2001) 11193–11198.

[6] C.A. Davis, L. Grate, M. Spingola, M. Ares Jr., Test of intron predictions reveals novel splice sites, alternatively spliced mRNAs and new introns in meiotically regulated genes of yeast, Nucleic Acids Res. 28 (2000) 1700–1706.

[7] M. Burset, I.A. Seledtsov, V.V. Solovyev, Analysis of canonical and non-canonical splice sites in mammalian genomes, Nucleic Acids Res. 28 (2000) 4364–4375.

[8] E. Bon, et al., Molecular evolution of eukaryotic genomes: hemiascomycetous yeast spliceosomal introns, Nucleic Acids Res. 31 (2003) 1121–1135.

[9] L. Huiet, N.H. Giles, The *qa* repressor gene of *Neurospora crassa*: wild-type and mutant nucleotide sequences, Proc. Natl. Acad. Sci. USA 83 (1986) 3381–3385.

[10] W.J. Kent, BLAT—the BLAST-like alignment tool, Genome Res. 12 (2002) 656–664.

[11] T. Farrer, A.B. Roller, W.J. Kent, A.M. Zahler, Analysis of the role of *Caenorhabditis elegans* GC-AG introns in regulated splicing, Nucleic Acids Res. 30 (2002) 3360–3367.

[12] I.J. Jackson, A reappraisal of non-consensus mRNA splice sites, Nucleic Acids Res. 19 (1991) 3795–3798.

[13] T.A. Thanaraj, F. Clark, Human GC-AG alternative intron isoforms with weak donor sites show enhanced consensus at acceptor exon positions, Nucleic Acids Res. 29 (2001) 2581–2593.

[14] D.J. Ebbole, et al., Gene discovery and gene expression in the rice blast fungus, Magnaporthe grisea: analysis of expressed sequence tags, Mol. Plant–Microbe Interact. 12 (2004) 1337–1347.

[15] F. Trail, J.R. Xu, P. San Miguel, R.G. Halgren, H.C. Kistler, Analysis of expressed sequence tags from *Gibberella zeae* (anamorph *Fusarium graminearum*), Fungal Genet. Biol. 38 (2003) 187–197.

[16] K. O'Donnell, T.J. Ward, D.M. Geiser, H.C. Kistler, T. Aoki, Genealogical concordance between the mating type locus and seven other nuclear genes supports formal recognition of nine phylogenetically distinct species within the *Fusarium graminearum* clade, Fungal Genet. Biol. 41 (2004) 600–623.

[17] A.A. Salamov, V.V. Solovyev, Ab initio gene finding in *Drosophila* genomic DNA, Genome Res. 10 (2000) 516–522.

[18] C. Burge, S. Karlin, Prediction of complete gene structures in human genomic DNA, J. Mol. Biol. 268 (1997) 78–94.

[19] T.J. Ward, et al., Intraspecific phylogeny and lineage group identification based on the *prfA* virulence gene cluster of *Listeria monocytogenes*, J. Bacteriol. 186 (2004) 4994–5002.

Web Site References

[1] *Magnaporthe* Sequencing Project, Ralph Dean, Fungal Genomics Laboratory at North Carolina State University: http://www.fungalgenomics.ncsu.edu.

[2] The Center for Genome Research: http://www.broad.mit.edu.

[3] *Fusarium graminearum* Sequencing Project, The Center for Genome Research: http://www.broad.mit.edu.

[4] *Neurospora crassa* Sequencing Project Assembly Version 3, The Center for Genome Research: http://www.broad.mit.edu.

[5] NCBI BLAST page: http://www.ncbi.nlm.nih.gov/BLAST/.